

# **MedHELM**: Holistic Evaluation of Large Language Models for Medical Tasks

Suhana Bedi<sup>1†</sup>, Hejie Cui<sup>1†</sup>, Miguel Fuentes<sup>1†</sup>, Alyssa Unell<sup>1†</sup>,  
Michael Wornow<sup>1</sup>, Juan M. Banda<sup>2</sup>, Nikesh Kotecha<sup>2</sup>,  
Timothy Keyes<sup>2</sup>, Yifan Mai<sup>3</sup>, Mert Oez<sup>4</sup>, Hao Qiu<sup>4</sup>, Shrey Jain<sup>4</sup>,  
Leonardo Schettini<sup>4</sup>, Mehr Kashyap<sup>1</sup>, Jason Alan Fries<sup>1</sup>,  
Akshay Swaminathan<sup>1</sup>, Philip Chung<sup>1</sup>, Fateme Nateghi<sup>1</sup>,  
Asad Aali<sup>1</sup>, Ashwin Nayak<sup>1</sup>, Shivam Vedak<sup>1</sup>, Sneha S. Jain<sup>1</sup>,  
Birju Patel<sup>1</sup>, Oluseyi Fayanju<sup>1</sup>, Shreya Shah<sup>1</sup>, Ethan Goh<sup>1</sup>,  
Dong-han Yao<sup>1</sup>, Brian Soetikno<sup>1</sup>, Eduardo Reis<sup>1</sup>,  
Sergios Gatidis<sup>1</sup>, Vasu Divi<sup>1</sup>, Robson Capasso<sup>1</sup>,  
Rachna Saralkar<sup>1</sup>, Chia-Chun Chiang<sup>1</sup>, Jenelle Jindal<sup>1</sup>,  
Tho Pham<sup>1</sup>, Faraz Ghoddusi<sup>1</sup>, Steven Lin<sup>1</sup>, Albert S. Chiou<sup>1</sup>,  
Christy Hong<sup>1</sup>, Mohana Roy<sup>1</sup>, Michael F. Gensheimer<sup>1</sup>,  
Hinesh Patel<sup>1</sup>, Kevin Schulman<sup>1</sup>, Dev Dash<sup>1</sup>, Danton Char<sup>1</sup>,  
Lance Downing<sup>1</sup>, Francois Grolleau<sup>1</sup>, Kameron Black<sup>1</sup>,  
Bethel Mieso<sup>1</sup>, Aydin Zahedivash<sup>1</sup>, Wen-wai Yim<sup>4</sup>,  
Harshita Sharma<sup>4</sup>, Tony Lee<sup>3</sup>, Hannah Kirsch<sup>2</sup>, Jennifer Lee<sup>2</sup>,  
Nerissa Ambers<sup>2</sup>, Carlene Lugtu<sup>2</sup>, Aditya Sharma<sup>2</sup>, Bilal Mawji<sup>2</sup>,  
Alex Alekseyev<sup>2</sup>, Vicky Zhou<sup>2</sup>, Vikas Kakkar<sup>2</sup>, Jarrod Helzer<sup>2</sup>,  
Anurang Revri<sup>2</sup>, Yair Bannett<sup>1</sup>, Roxana Daneshjou<sup>1</sup>,  
Jonathan Chen<sup>1</sup>, Emily Alsentzer<sup>1</sup>, Keith Morse<sup>1</sup>, Nirmal Ravi<sup>1</sup>,  
Nima Aghaeepour<sup>1</sup>, Vanessa Kennedy<sup>1</sup>, Akshay Chaudhari<sup>1</sup>,  
Thomas Wang<sup>1,2</sup>, Sanmi Koyejo<sup>3, 5</sup>, Matthew P. Lungren<sup>1,4</sup>,  
Eric Horvitz<sup>4, 5</sup>, Percy Liang<sup>3</sup>, Mike Pfeffer<sup>2</sup>, Nigam H. Shah<sup>1,2\*</sup>

<sup>1</sup>Stanford University School of Medicine, Stanford, CA, USA.

<sup>2</sup>Stanford Health Care, Palo Alto, CA, USA.

<sup>3</sup>Center for Research on Foundation Models (CRFM) & Department of Computer Science, Stanford University, CA, USA.

<sup>4</sup>Microsoft Corporation, Redmond, WA, USA.

<sup>5</sup>Stanford Institute for Human-Centered AI, CA, USA.

\*Corresponding author(s). E-mail(s): [nigam@stanford.edu](mailto:nigam@stanford.edu);

†These authors contributed equally to this work and are listed in alphabetical order.

### Abstract

While large language models (LLMs) achieve near-perfect scores on medical licensing exams, these evaluations inadequately reflect the complexity and diversity of real-world clinical practice. We introduce MedHELM, an extensible evaluation framework for assessing LLM performance for medical tasks with three key contributions. First, we present a clinician-validated taxonomy spanning five categories, 22 subcategories, and 121 tasks developed with 29 clinicians. Second, we develop a comprehensive benchmark suite comprising 35 benchmarks (17 existing, 18 newly formulated) providing complete coverage of all categories and subcategories in the taxonomy. Third, we conduct a systematic comparison of LLMs with improved evaluation methods (using an LLM-jury) and a cost-performance analysis. Evaluation of nine frontier LLMs, using the 35 benchmarks, revealed significant performance variation. Advanced reasoning models (DeepSeek R1: 66% win-rate; o3-mini: 64% win-rate) demonstrated superior performance, though Claude 3.5 Sonnet achieved comparable results at 40% lower estimated computational cost. On a normalized accuracy scale (0-1), most models performed strongly in Clinical Note Generation (0.74-0.85) and Patient Communication & Education (0.76-0.89), moderately in Medical Research Assistance (0.65-0.75) and Clinical Decision Support (0.61-0.76), and lower in Administration & Workflow (0.53-0.63). Our LLM-jury evaluation method achieved good agreement with clinician ratings ( $ICC = 0.47$ ), surpassing both average clinician-clinician agreement ( $ICC = 0.43$ ) and automated baselines, including ROUGE-L (0.36) and BERTScore-F1 (0.44). Claude 3.5 Sonnet achieved comparable performance to top models at lower estimated cost. These findings highlight the importance of real-world, task-specific evaluation for medical use of LLMs and provides an open source framework to enable this.

**Keywords:** large language models, evaluation, medicine, benchmark, taxonomy

## 1 Introduction

Large Language Models (LLMs) have shown impressive performance on medical knowledge benchmarks, achieving  $\sim 99\%$  accuracy on standardized exams like MedQA [1]. This has sparked interest in deploying them in healthcare settings: supporting clinical decision-making such as diagnosis and treatment [2], optimizing clinical workflows including documentation and scheduling [3], and enhancing patient education and communication [4].

However, there is a large gap between performance on medical knowledge benchmarks and readiness for real-world deployment due to three key limitations in these existing benchmarks [5]: (1) *Questions do not match real-world settings* – Existing

benchmarks rely on synthetic vignettes or narrowly-scoped exam questions, failing to capture key aspects of real diagnostic processes such as extracting relevant details from patient records [6, 7]. (2) *Limited use of real-world data* – Only 5% of LLM evaluations use real-world electronic health record (EHR) data [8]. EHRs contain ambiguities, inconsistencies, and domain-specific shorthand that synthetic data cannot replicate. (3) *Limited task diversity* – Around 64% of LLM evaluations in healthcare focus only on medical licensing exams and diagnostic tasks [8], ignoring essential hospital operations such as administrative tasks (e.g., generating prior authorization letters, identifying billing codes), clinical documentation (e.g., writing progress notes or discharge instructions), and patient communication (e.g., asynchronous messaging through electronic patient portals) [9].

Recent work on HealthBench [10] has advanced the evaluation of LLMs in medicine by scoring 5000 single-turn, free-text dialogues in which the model acts independently, much like a direct-to-patient advice line, without follow-up questions or human oversight. Its physician-authored rubrics reward exhaustive, risk-averse responses that maximize safety and completeness, offering a valuable stress test for fully autonomous chatbots. However, this design does not capture the iterative, context-aware interactions clinicians expect from an assistive co-pilot, nor does it assess performance on structured tasks that dominate everyday workflows (e.g., order review, note generation, literature summarization).

To address these limitations, we introduce **MedHELM** (Holistic Evaluation of Large Language Models for Medical Tasks), an extensible evaluation framework for assessing LLM performance in completing medical tasks (Figure 1). Inspired by the HELM project’s standardized cross-domain evaluations [11], using **MedHELM** we evaluate nine LLMs using 35 distinct benchmarks covering all 22 subcategories of medical tasks, focusing on clinicians’ day-to-day activities beyond just taking licensing exams. We assess performance using benchmark-appropriate metrics (i.e., exact match for closed-ended benchmarks, LLM-jury where three LLMs evaluate responses using tailored rubrics for open-ended benchmarks with demonstrated agreement to clinician ratings) as well as estimated computational cost to provide practical deployment insights. Our primary contributions are:

1. **Clinician-validated taxonomy:** A five-category, 22-subcategory, 121-task taxonomy developed with 29 clinicians. Clinicians achieved a 96.7% agreement rate when mapping subcategories to appropriate categories, validating the clear and discrete nature of the taxonomy. We present the complete taxonomy in the Results and Appendix sections.
2. **A benchmark suite with full taxonomy coverage:** A collection of 35 benchmarks spanning all 22 subcategories of medical tasks. This includes 17 existing benchmarks, five re-formulated benchmarks based on existing datasets, and 13 new benchmarks.<sup>1</sup>
3. **Comparative evaluation of models along with cost-performance analysis:** A systematic evaluation shows that reasoning models achieve the highest overall performance.

---

<sup>1</sup>For privacy and regulatory compliance, as well as to prevent inclusion in LLM training data, 14 datasets are not publicly released.

The **MedHELM** framework addresses a critical need in the testing and evaluation of AI for medical use by providing consistent, real-world evaluation standards for medical (and healthcare) applications of LLMs. This framework benefits three key stakeholder groups: (1) Healthcare systems evaluating LLMs for specific tasks, (2) AI developers identifying performance gaps across medical tasks, and (3) Researchers developing methods to reproducibly measure LLM capabilities on medical tasks. To foster collaborative improvement of such AI evaluation, we provide an openly accessible **leaderboard** <sup>2</sup> with current benchmarking results and share the **codebase** <sup>3</sup>, with **documentation** <sup>4</sup> for contributing new datasets, evaluation metrics, and benchmarking custom models. By standardizing terminology and evaluation methods across the task taxonomy, **MedHELM** establishes a foundation for reproducible and real-world assessment of AI capabilities in medicine.

## 2 Results

### 2.1 Clinician Validation of the Taxonomy

In a structured review process, 29 clinicians evaluated the initial taxonomy comprising five categories, 21 subcategories, and 98 tasks. When asked to assign each subcategory to its appropriate top-level category, clinicians correctly matched 96.7% of subcategories to their intended categories. The clinicians rated the comprehensiveness of the proposed tasks at a mean of 4.21/5 ( $n = 29$ ) and provided 107 comments with suggestions for improvement. Based on this feedback, we refined task definitions and expanded the taxonomy to its final form: five categories, 22 subcategories, and 121 tasks. An overview of the final taxonomy is shown in Figure 2, with the complete list provided in Appendix A.

### 2.2 Overview of the Benchmark Suite

Our 35 benchmarks span all 22 subcategories, providing full coverage over categories and subcategories in our taxonomy (Table C3). The benchmark suite comprises 17 *existing* benchmarks, five *re-formulated* benchmarks derived from previously unevaluated medical datasets, and 13 *new* benchmarks, out of which 12 are EHR-based. The suite includes 13 open-ended benchmarks (requiring free-text generation) and 22 closed-ended benchmarks (with predefined answer choices). Access levels are designated as 14 public, seven gated (i.e. requiring approval), and 14 private.

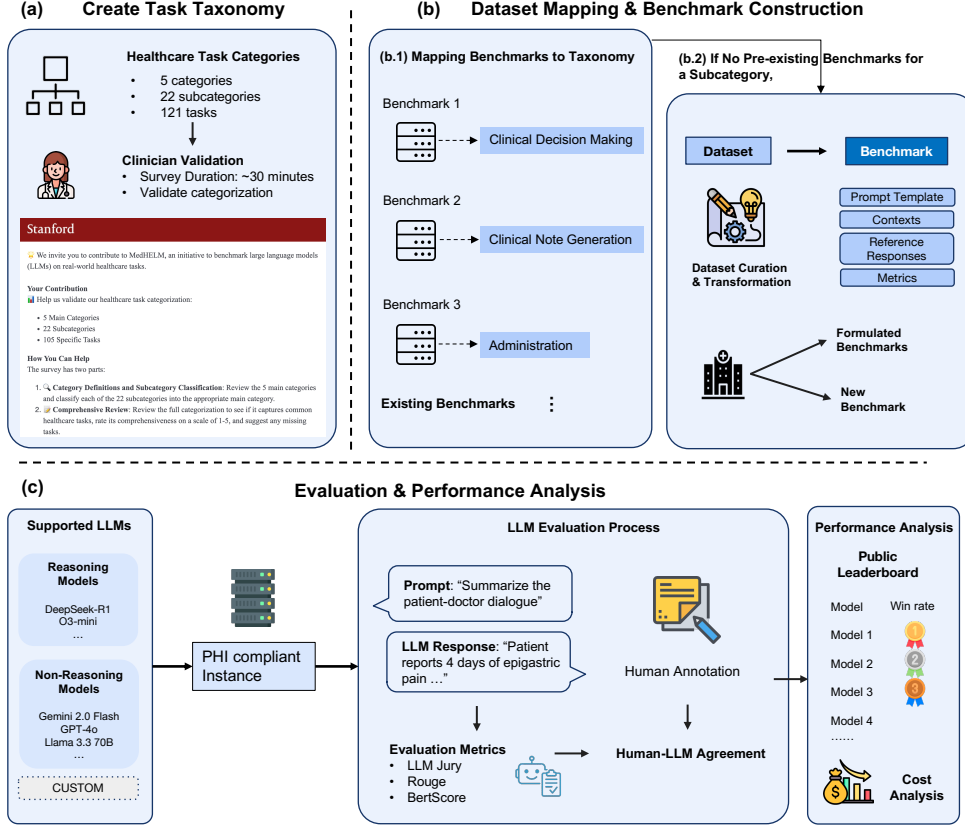
*Clinical Decision Support* is the most represented category with ten benchmarks, followed by *Patient Communication* (eight), *Clinical Note Generation* and *Medical Research Assistance* (six), and *Administration & Workflow* (five). The distribution of benchmarks across subcategories is uneven, with 15 subcategories containing a single benchmark and the remaining seven subcategories containing between two and five benchmarks each.

---

<sup>2</sup><https://crfm.stanford.edu/helm/medhelm/v2.0.0/>

<sup>3</sup><https://github.com/stanford-crfm/helm>

<sup>4</sup><https://crfm-helm.readthedocs.io/en/latest/medhelm/>



**Fig. 1** This figure illustrates: (a) a clinician-validated taxonomy organizing 121 medical tasks into five categories and 22 subcategories; (b) a suite of benchmarks that map existing benchmarks to this taxonomy and introduces new benchmarks for complete coverage; and (c) an evaluation comparing reasoning and non-reasoning LLMs, with model rankings, LLM jury based evaluation of open-ended benchmarks, and cost-performance analysis

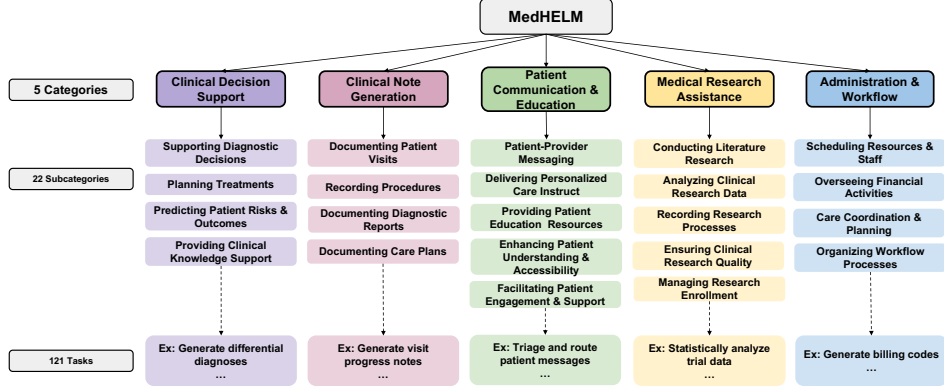
## 2.3 Model Evaluation & Cost-Performance Analysis

### 2.3.1 Overall Performance

#### *Pairwise Win-Rate and Average Scores*

Table 1 compares models using win-rate and macro-average performance metrics (defined in the table caption). DeepSeek R1 performed best, winning 66% of head-to-head comparisons with a macro-average of 0.75 and low win standard deviation (0.10). o3-mini followed with a 64% win rate and the highest macro-average (0.77), driven by strong performance in benchmarks in the clinical decision support category.

The Claude models achieved 63-64% win rates and identical macro-averages (0.73). GPT-4o achieved a 57% win rate, while Gemini 2.0 Flash (42%) and GPT-4o mini (39%) performed lower. Open-source Llama 3.3 Instruct achieved a 30% win rate.



**Fig. 2** Overview of the final taxonomy comprising five main categories and 22 subcategories.

Gemini 1.5 Pro ranked lowest with 24% wins but had the lowest win standard deviation (0.08), showing the most consistent competitive performance.

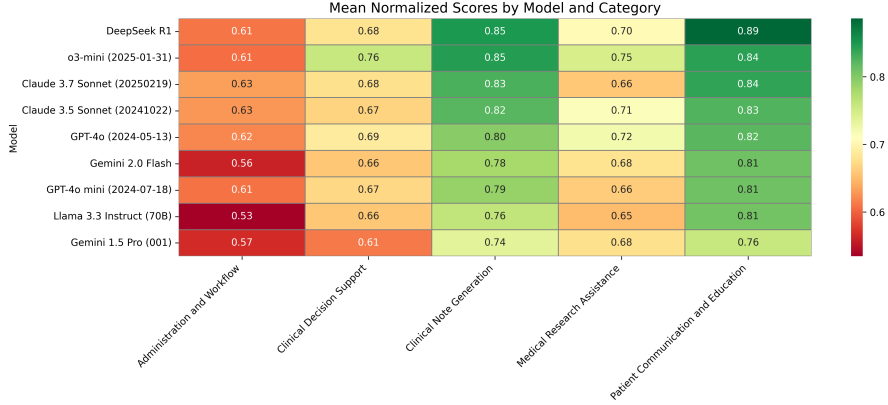
Model (snapshot)	Win-rate $\uparrow$	Win SD $\downarrow$	Macro-avg $\uparrow$	SD $\downarrow$
DeepSeek R1	<b>0.66</b>	0.10	0.75	0.22
o3-mini (2025-01-31)	0.64	0.16	<b>0.77</b>	<b>0.18</b>
Claude 3.7 Sonnet (20250219)	0.64	0.13	0.73	0.21
Claude 3.5 Sonnet (20241022)	0.63	0.14	0.73	0.21
GPT-4o (2024-05-13)	0.57	0.17	0.73	0.18
Gemini 2.0 Flash	0.42	0.17	0.70	0.21
GPT-4o mini (2024-07-18)	0.39	0.18	0.71	0.20
Llama 3.3 Instruct (70B)	0.30	0.13	0.69	0.22
Gemini 1.5 Pro (001)	0.24	<b>0.08</b>	0.67	0.21

**Table 1** Comparison of performance of frontier models across 35 MedHELM benchmarks, sorted by descending win-rate. **Bold** indicates the best value in each column. Win-rate represents the proportion of pairwise comparisons where each model achieved superior performance across all 35 benchmarks (possible range: 0-1). Win standard deviation (SD) measures how consistently a model wins (lower values = more consistent). Macro-avg is the average performance score across all 35 benchmarks. SD shows how much performance varies across different benchmarks (lower values = more consistent across benchmarks).

### Performance by Benchmark

We present every model’s normalized score from each of the 35 benchmarks as a heatmap in Figure 3, where darker green indicates higher performance. Models perform worse on benchmarks such as MedCalc-Bench (calculating medical values from patient notes), EHRSQL (generating SQL queries from natural language instructions for clinical research, originally intended as a code generation dataset), and MIMIC-IV Billing Code (assigning ICD-10 codes to clinical notes). The best performance is seen for the NoteExtract benchmark (extracting specific information from clinical notes). Broader category-level trends are described below.





**Fig. 4** Mean normalized scores (0-1 scale) across the five categories for all evaluated models. Darker green represents higher scores. Models are ordered by mean win rate from top (highest) to bottom (lowest), while categories are arranged left to right.

*Communication & Education* (0.81) but shows the lowest score in *Administration & Workflow* (0.53), indicating areas for future improvement.

### 2.3.2 Evaluation of Open-Ended Benchmarks

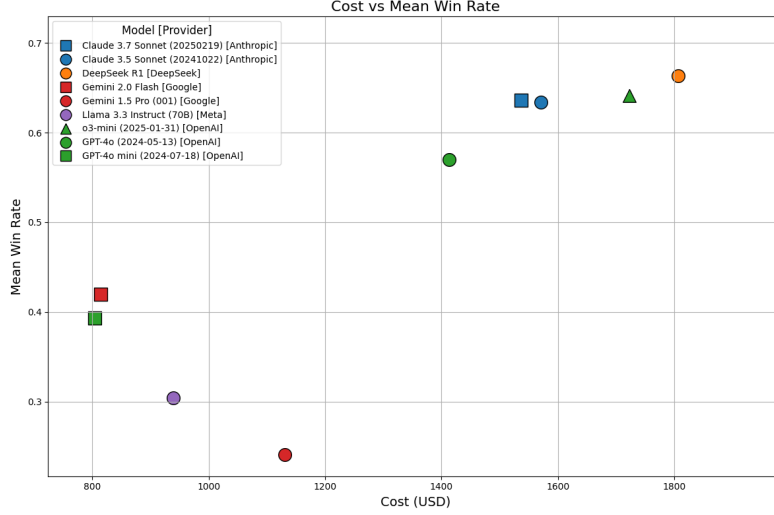
For our 13 open-ended benchmarks, we implemented an LLM-jury evaluation approach. To assess this method’s validity, we collected independent clinician ratings on a subset of model outputs. We used 31 instances from ACI-Bench and 25 from MEDIQA-QA to compare clinician-assigned scores to the jury’s aggregated ratings as described in the methods. Table 2 reports intraclass correlation coefficient (ICC(3,k)) values (after rater-wise  $z$ -scoring) for the LLM-jury versus clinicians, alongside ROUGE-L (text overlap metric) and BERTScore-F (semantic similarity metric) baselines, and the average clinician–clinician agreement.

Overall, the LLM-jury achieves an ICC of 0.47, which beats the average clinician–clinician agreement (ICC = 0.43) and automated baselines including ROUGE-L (0.36) and BERTScore-F (0.44). These results demonstrate that our LLM-jury mirrors clinician judgment better than standard lexical metrics, establishing its validity as a stand-in for clinician raters.

Benchmark	LLM	ROUGE-L	BERTScore-F	Clinician
Combined	0.474 (0.100, 0.690)	0.361 (0, 0.630)	0.441 (0.050, 0.670)	0.426 (0.295, 0.585)
ACI-Bench	0.305 (0, 0.670)	0.445 (0, 0.730)	0.250 (0, 0.640)	0.458 (0.201, 0.945)
MEDIQA	0.625 (0.150, 0.830)	0.343 (0, 0.710)	0.668 (0.250, 0.850)	0.520 (0.500, 0.534)

**Table 2 Agreement of LLM-jury and of automated metrics with clinician ratings** The table entries are ICC(3,k) coefficients after  $z$ -scoring within rater (ICC3k- $z$ ); 95 % confidence intervals are shown in parentheses. Higher ICC indicates better agreement with clinician ratings. The last column gives the average clinician–clinician agreement for each dataset.





**Fig. 5** Scatter plot of mean win-rate (y-axis) versus estimated computational cost (x-axis) for each of the nine models across 35 benchmarks. Each point represents a model, with the position indicating the relationship between performance (y-axis) and total cost of evaluation, including benchmark runs and evaluation by LLM-jury (x-axis). Costs represent upper-bound estimates based on maximum output token usage.

### 2.3.3 Cost–Performance Analysis

We estimated the cost (USD) of evaluating each model based on publicly listed pricing as of 05/12/2025, using the total input tokens and maximum output tokens consumed during benchmark runs and LLM-jury evaluation. These costs are an upper-bound estimate, since models may generate fewer tokens than the maximum allowed output. We plotted the mean win-rate against the cost (Figure 5, with a detailed breakdown in Table 3, and inference costs in Appendix G). As expected, non-reasoning models—GPT-4o mini (\$805) and Gemini 2.0 Flash (\$815)—incurred the lowest costs and achieved win-rates of 0.39 and 0.42, respectively. Open-source Llama 3.3 Instruct (\$940) had a 0.30 win-rate, while Gemini 1.5 Pro (\$1,130) reached 0.24. Reasoning models incurred higher costs, DeepSeek R1 (\$1,806) and o3-mini (\$1,722), with win-rates of 0.66 and 0.64, respectively. Claude 3.5 Sonnet (\$1,571) and Claude 3.7 Sonnet (\$1,537) provide a good cost-performance balance, achieving  $\sim 0.63$  win-rate at reduced costs.

Model	Model Creator	Window Size	Access	Benchmark Tokens	Jury Tokens	Benchmark Cost	Jury Cost	Total Cost
Claude 3.5 Sonnet (20241022)	Anthropic	200,000	Closed	245,958,343	45,868,483	\$778.67	\$792.21	\$1,570.88
Claude 3.7 Sonnet (20250219)	Anthropic	200,000	Closed	242,510,517	42,326,449	\$768.26	\$768.38	\$1,536.64
Gemini 1.5 Pro (001)	Google	1,000,000	Closed	277,273,571	42,864,352	\$359.33	\$771.73	\$1,131.06
Gemini 2.0 Flash	Google	1,000,000	Closed	276,777,154	42,864,352	\$43.04	\$771.73	\$814.77
GPT-4o (2024-05-13)	OpenAI	128,000	Closed	248,731,118	41,929,516	\$647.28	\$765.91	\$1,413.19
GPT-4o mini (2024-07-18)	OpenAI	128,000	Closed	248,731,118	41,929,516	\$38.83	\$765.91	\$804.74
Llama 3.3 Instruct (70B)	Meta AI	128,000	Open	242,895,608	42,121,933	\$172.45	\$767.13	\$939.58
DeepSeek R1	DeepSeek	128,000	Open	334,133,126	68,364,208	\$848.21	\$957.96	\$1,806.17
o3-mini (2025-01-31)	OpenAI	128,000	Closed	337,967,189	68,617,978	\$762.50	\$959.54	\$1,722.04

**Table 3** Comparison of LLMs by architecture, access type, and token usage metrics across MedHELM. Benchmark tokens denote total input/output tokens in completing the medical task represented by the benchmark; jury tokens are tokens used for open-ended evaluations via the LLM-jury. The total cost reflects the estimated per-model expense of running a MedHELM evaluation across 35 benchmarks and represents an upper bound based on maximum output token usage. For nine models and 35 benchmarks in the current MedHELM suite, one update to the leaderboard is estimated to cost 11,739.07.

### 3 Discussion

We present a framework for assessing LLM performance for real-world medical tasks. Our clinician-validated taxonomy provides a structure for summarizing models’ strengths and limitations across medical tasks. High clinician agreement (96.7%) in assigning subcategories to categories suggests the taxonomy effectively captures how healthcare professionals conceptualize their work. Our benchmark suite reveals nuances in model capabilities that are not seen with current medical knowledge benchmarks alone. The higher performance in communication tasks than administrative ones may stem from administrative workflows using data not seen during training, warranting caution in healthcare AI implementation for back-office tasks without quantifying task-specific performance. This framework addresses the primary limitations of current benchmarks, such as the lack of using real-world data, use of evaluation setups that do not match real-world settings, and limited task diversity. The cost-performance trade-off shows that while reasoning models have superior performance, their substantially higher costs may not justify their deployment for all tasks. In a resource-constrained setting, models such as Claude 3.5 Sonnet offer a balance, achieving a win-rate of 0.63 at a lower cost. Finally, our LLM-jury based approach addresses a critical gap in current evaluation approaches. By beating clinician-clinician agreement, this approach enables scalable evaluation of open-ended model outputs without requiring extensive clinician time, a scarce and expensive resource.

Several limitations remain. While our LLM-jury approach was validated on only two benchmarks, expanding clinician annotations across more benchmarks would strengthen the clinician–LLM agreement estimates. In addition, the uneven distribution of benchmarks across subcategories (15 of 22 contain only one benchmark) limits our ability to draw robust performance conclusions in underrepresented areas. Moreover, our current rubrics operate at the benchmark level, but instance-level rubrics could provide better evaluation, particularly for subjective or context-dependent medical tasks where gold standard responses may not exist. Such approaches would further scale LLM-jury evaluation beyond reliance on gold standard responses [12]. Administration & Workflow emerged as the weakest performance area for all models. Understanding the underlying causes of this poor performance, whether stemming from training data

limitations, task complexity, or distributional shifts, is essential for safe deployment in healthcare operations.

In conclusion, **MedHELM** provides a comprehensive framework for assessing LLM performance across real-world medical tasks through our clinician-validated taxonomy spanning five categories, 22 subcategories, and 121 tasks. Our benchmark suite of 35 datasets reveals that most models perform best in Clinical Note Generation (0.74-0.85) and Patient Communication & Education (0.76-0.89), moderately in Medical Research Assistance (0.65-0.75) and Clinical Decision Support (0.61-0.76), and worst in Administration & Workflow (0.53-0.63). Reasoning models DeepSeek R1 and o3-mini led overall with win-rates of 0.66 and 0.64, respectively, though Claude models offer competitive performance (win-rate of  $\sim 0.63$ ) at lower computational cost. Through our public leaderboard<sup>5</sup> and shared codebase<sup>6</sup>, **MedHELM** establishes infrastructure for ongoing collaborative assessment as models evolve, advancing medical AI evaluation that better reflects the complexity of real-world medical practice.

## 4 Methods

**Motivation and related work.** Most LLM evaluations in medicine still rely on closed-form question-answering over exam-style datasets such as MEDQA and MEDMCQA, with only  $\sim 5\%$  incorporating real EHR data and very few addressing free-text generation tasks or cost-aware metrics [7, 13, 14]. Large-scale Natural Language Processing (NLP) meta-benchmarks (HELM, BIG-BENCH) demonstrate the value of task diversity and multi-metric scoring [11, 15], while biomedical efforts like ClinicBench [16] or MMedBench [17] each advance a single dimension (e.g. multimodality or cost-aware metrics) without clinician-validated scope or extensible tooling. Recent work on HEALTHBENCH [10] addresses some of these limitations by incorporating physician-developed rubrics for 5,000 health conversations, but lacks a comprehensive taxonomy of medical tasks and is done with synthetic data. Our goal is to close these gaps by co-designing, with clinicians, a taxonomy-guided benchmark suite that (i) spans the full spectrum of real medical work, (ii) uses both public and private medical record data, and (iii) uses evaluation protocols that align with clinician judgment.

### 4.1 Development of the Taxonomy of Tasks

Early attempts (BIGBIO, ClinicBench) at creating a taxonomy of tasks in medical settings either harmonize datasets into broad categories without clinician input or collapse heterogeneous skills under a single “generation” label [16, 18]. To ensure our benchmarks accurately reflect the complexity of medical practice, we developed a taxonomy that mirrors how clinicians conceptualize their daily work. By defining a hierarchical structure, we guarantee that each benchmark maps to a concrete medical activity. Our taxonomy consists of three levels:

- **Category:** A broad domain of medical activity (e.g., *Clinical Decision Support*).

---

<sup>5</sup><https://crfm.stanford.edu/helm/medhelm/latest/>

<sup>6</sup><https://github.com/stanford-crfm/helm>

- **Subcategory:** A group of related tasks within a Category (e.g., *Supporting Diagnostic Decisions*).
- **Task:** A discrete action taken during the delivery of medical care (e.g., *Generate differential diagnoses*).

This structure enables systematic coverage of the medical care landscape while maintaining clear boundaries between distinct activities that care providers perform.

#### 4.1.1 Initial drafting

We based our taxonomy on tasks identified in a JAMA review [8]. Working with a clinician (MK), we reorganized these tasks into functional themes that reflect real-world activities, resulting in 98 distinct tasks organized into 21 subcategories within five categories:

1. Clinical Decision Support
2. Clinical Documentation
3. Patient Communication & Education
4. Medical Research Assistance
5. Administration & Workflow

Two principles guided our taxonomy development:

- **Medical relevance:** Each task maps directly to actions routinely performed by care providers.
- **Clear boundaries:** Categories and subcategories were defined to minimize overlap while preserving meaningful functional distinctions.

#### 4.1.2 Validation

To validate our initial taxonomy, we designed a two-part survey completed by 29 practicing clinicians across 14 medical specialties. More information on participating clinicians can be found in Appendix B. The survey assessed both categorical organization and real-world relevance:

In the first section, clinicians assigned each of our 21 subcategories to one of the 5 main categories. This exercise tested whether our taxonomy structure matched how clinicians naturally organize medical tasks.

In the second section, clinicians evaluated the comprehensiveness of our taxonomy on a 5-point scale, where a score of 5 means that our categorization covered all routine medical tasks and a score of 1 means it covered very few. They also provided feedback through an open dialogue box where they could suggest missing tasks and recommend terminology improvements. This systematic validation approach evaluated both the taxonomy’s organizational logic and its comprehensiveness in representing actual medical tasks.

Based on the comments, we refined definitions and expanded the taxonomy to have 5 categories, 22 subcategories, and 121 tasks.

## 4.2 Construction of the Benchmark Suite

### 4.2.1 Curation of datasets

To construct a comprehensive suite of 35 benchmarks spanning our taxonomy, we employed a three-tiered dataset curation strategy:

1. **Existing benchmarks:** We incorporated existing benchmarks from public or gated sources (e.g., MedQA, MIMIC-IV Billing Code, ACI-Bench) to ensure broad subcategory coverage.
2. **Reformulated benchmarks:** We transformed previously unevaluated medical data collections into "reformulated benchmarks" by applying standardized prompt templates and specifying evaluation metrics. This approach addressed subcategories where datasets existed but lacked LLM-ready evaluation benchmarks.
3. **New benchmarks:** To address the under-representation of *Administration & Workflow*, we partnered with Stanford Healthcare to curate private datasets for tasks that are routinely done in health systems but for which benchmark datasets do not exist (e.g. referral triage, scheduling).

Each benchmark is labeled by *source type* (existing / reformulated / new) and *access level* (public / gated / private), with the provenance documented in the **MedHELM** repository.

### 4.2.2 Specification of Prompts and Metrics

To transform each curated dataset into a **MedHELM** benchmark, we defined three components for every item in the dataset:

- **Context:** the raw input presented to the LLM (e.g. a clinical note, patient message, or structured EHR record).
- **Prompt:** a standardized instruction template to elicit consistent, task-appropriate responses (e.g. "Answer in 2–3 sentences" for open-ended summaries, or MCQ framing for multiple-choice questions).
- **Evaluation Metric:** a pre-specified scoring method matched to the task type:
  - *Exact-match accuracy* for single-token or numeric outputs (e.g. selecting the correct option in MedQA).
  - *Micro-F1* for multi-label classification tasks (e.g. ICD-10 code assignment).
  - *LLM-jury ensemble* for open-ended text generation: we use a three-model Likert-scale protocol assessing medical accuracy, completeness and clarity, and secondary metrics, ROUGE and BERTScore, to capture lexical and semantic overlap.
- **Gold Standard Response (Optional):** the reference output (numeric result, classification label, or sample text) against which the model’s response is scored (e.g., "4" in response to "What’s a patient’s HAS-BLED score?"). While most benchmarks include a gold standard, our framework accommodates benchmarks without one, such as NoteExtract, providing flexibility for future evaluation needs.

Table 4 illustrates these components for the HeadQA benchmark (subcategory: Providing Clinical Knowledge Support) and the MedDialog benchmark (subcategory: Patient-Provider Messaging).

We found quality issues in some of the gold standard responses for reformulated benchmarks. For example, gold standard responses in the MIMIC-RRS benchmark occasionally contained information from a patient’s EHR that was not passed into the model’s context. To assess the impact of these low-quality gold standard responses, we conducted a sensitivity analysis by filtering "problematic" gold standard responses using an LLM judge and recalculating metrics (Appendix D). Model rankings remained unchanged, as instances with "problematic" and "non-problematic" gold standard responses received similar jury scores. This stability exists because our LLM-jury (prompt in Appendix F) uses gold standard responses only when needed.

Benchmark	Context	Prompt	Evaluation Metric
HeadQA	<p><i>Question:</i> In the stools of a 6-year-old boy with diarrhea and abdominal pain, a packet of parasite eggs surrounded by a transparent membrane is observed. Which treatment do you consider most appropriate?</p> <p>A. Praziquantel B. Azithromycin C. Metronidazole D. Albendazole</p>	<p>You are a highly knowledgeable AI assistant specializing in biomedical sciences. Your task is to answer multiple-choice questions accurately based on the options provided. Select the correct answer by outputting only the letter corresponding to your choice (A, B, C, or D).</p> <p><b>Question:</b> {Context} <b>Answer:</b></p>	<p>Exact Match <b>Answer: D</b></p>
MedDialog	<p><i>Patient:</i> I have had this pain in my left arm down to almost my wrist. It first started where I would have a sharp pain between my armpit and boob. It doesn't hurt all the time but has become more often lately. It's an aching throbbing feeling but also a numbing weak feeling to where I lose my strength in my arm. After a while it will pass but I don't know if it's something serious. Need help.</p> <p><i>Doctor:</i> Hi, take this seriously, it could be a lump compressing on your nerves, or any infection. You need an examination by a doctor to exactly judge what your symptoms mean and what is their cause. I recommend you consult your doctor at the earliest.</p>	<p>Generate a one sentence summary of this patient-doctor conversation.</p>	<p>LLM Jury <b>Accuracy: 4.88</b></p>

**Table 4** Case study of HeadQA and MedDialog, two benchmarks under MedHELM

## 4.3 Model Evaluation and Cost-Performance Analysis

### 4.3.1 Model Selection and Inference Pipeline

We evaluated 9 state-of-the-art LLMs (Table 1) under a uniform prompting and decoding regimen. All models were queried via their respective APIs or local endpoints

with sampling temperature set to 0 for deterministic outputs. All experiments were conducted on a PHI-compliant shared cluster maintaining full HIPAA compliance.

### 4.3.2 Performance Metrics

To quantify task performance, we computed:

- **Pairwise win-rate:** For each of the 35 benchmarks, we compared each model against every other; a “win” is assigned if a model’s normalized score  $\geq$  its rival’s. We then averaged wins over all pairings.
- **Macro-average score:** The overall performance score calculated by averaging results across all 35 benchmarks, with each benchmark weighted equally regardless of size (0–1 scale).

### 4.3.3 Evaluation of Open-Ended Benchmarks

For open-ended benchmarks, early pipelines used a single “LLM-as-Judge” to score outputs [19, 20], but high variance and bias spurred the “LLM-as-Jury” paradigm, aggregating  $k$  independent judgments for closer expert agreement [21, 22]. Extensions like G-Eval introduce chain-of-thought (CoT) self-critique per juror, and tools such as SelfCheckGPT and FActScore target hallucination and factuality [23–25]. MedHELM adopts a three-member jury without CoT to balance reliability and runtime.

#### *Evaluation using LLM-Jury*

We selected a three-member jury based on prior research demonstrating that odd-numbered panels reduce tie scenarios while maintaining reliability [26]. The jury composition (GPT-4o, Claude 3.7 Sonnet, LLaMA 3.3 70B) was chosen to represent diverse model architectures and training approaches, minimizing systematic bias from any single provider. We prompted each judge to score the model-generated responses on a 1–5 Likert scale according to three axes adopted from [27]:

- **Accuracy:** Factual correctness and adherence to medical guidelines.
- **Completeness:** Thoroughness in addressing all aspects of the query.
- **Clarity:** Organization, readability, and easy to understand language.

For the NoteExtract benchmark, which requires restructuring free-form clinical care plans into a specified format without a gold standard response, we modified our evaluation approach. We replaced the “completeness” criterion with “structure,” which assessed whether model responses properly reorganized the input text according to the requested format. The final LLM-jury score for each response is the mean of all nine ratings ( $3 \text{ judges} \times 3 \text{ axes}$ ).

#### *Clinician Rating*

To validate the LLM-jury approach, we collected clinician ratings on a subset of two open-ended benchmarks (MEDIQA and ACI-Bench). We selected these benchmarks because they were publicly available (facilitating annotation) and represent different categories (Patient Communication and Education and Clinical Note Generation, respectively). 20 clinicians across various specialties each scored a subset of responses on the same three axes, with at least two clinicians per instance.

### *Clinician–LLM Agreement Metrics*

We assessed agreement between the LLM-jury and clinicians via two intraclass correlation coefficients (ICC), after applying z-score normalization to each rater’s composite (mean of the three axis ratings) to remove scale biases:

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i},$$

where  $x_{ij}$  is rater  $i$ ’s composite on instance  $j$ .

1. **ICC(3,k)<sub>z</sub>**: We treat the clinicians’ mean and LLM jury mean as two fixed raters in a two-way mixed-effects model, thus, we compute

$$\text{ICC}(3, k) = \frac{MS_{\text{cases}} - MS_{\varepsilon}}{MS_{\text{cases}}}, \quad k = 2,$$

to quantify absolute agreement.

2. **Average Clinician–Clinician ICC<sub>z</sub>**: For every pair of clinicians who scored at least two common instances, we z-normalize each rater’s composite scores and compute ICC(3,k=2). We then average these pairwise ICCs (and bootstrap a 95 % CI) to give a representative inter-clinician agreement baseline.

Together, these metrics measure (i) the fidelity of LLM-jury to expert clinician rating and (ii) whether LLM–Clinician alignment approaches inter-clinician agreement.

ROUGE and BERTScore were also computed for all open-ended benchmarks with gold standard responses but used only as secondary metrics due to their limited alignment with clinical judgment.

### **4.3.4 Cost–Performance Analysis**

We tracked token usage for both benchmark prompts and jury evaluations across all models by counting input tokens and assuming the maximum allowed output tokens. This approach was necessary because, for o3-mini, it is not possible to fully observe the token usage as the thinking tokens don’t appear in the final output. As a result, our cost estimates represent an upper bound. Using published per-1M-token pricing as of 05/12/2025, we estimated the total cost per model. Cost–performance trade-offs were visualized by plotting mean win-rate against estimated total evaluation cost, highlighting models that optimize accuracy per dollar spent.

By integrating clinician-validated taxonomy design, a balanced mix of public and private datasets, and LLM-jury evaluation, our benchmark addresses the three structural gaps identified in earlier work—task diversity, medical grounding, and metric fidelity—while providing transparent cost accounting for real-world deployment decisions.



# Appendix A Task Taxonomy

## Table of Contents

1. Overview .....	1
2. Clinical Decision Support .....	2
• Definition	
• Subcategories and Tasks	
3. Clinical Note Generation .....	3
• Definition	
• Subcategories and Tasks	
4. Patient Communication and Education .....	5
• Definition	
• Subcategories and Tasks	
5. Medical Research Assistance .....	7
• Definition	
• Subcategories and Tasks	
6. Administration and Workflow .....	9
• Definition	
• Subcategories and Tasks	

## Overview

- **Total Categories:** 5
- **Entities per Category:**
  - **Definition:** A description of the category.
  - **Inclusion criteria:** Guidelines determining which subcategories (and tasks) belong in the category.
  - **Task performer:** Individuals responsible for executing tasks under the category.
  - **Subcategories:** Each category contains multiple subcategories.
  - **Tasks:** Each subcategory comprises several tasks.

### Category

- Definition
- Inclusion criteria
- Task performer
- Sub category
  - Task
  - Task
  - Task
- Sub category
  - Task
  - Task
  - Task

## 1. Clinical Decision Support

**Definition:** Analyzing patient-specific data to provide evidence-based recommendations to clinicians.

**Inclusion criteria:** Actions generating actionable insights based on patient data and medical evidence.

**Task performer:** Healthcare professionals (clinicians, nurses, and other practitioners).

### Subcategories and Tasks:

- **Supporting Diagnostic Decisions**
  - Recognize disease patterns from symptoms/vitals
  - Interpret diagnostic tests (ECG, spirometry)
  - Generate follow-up questions
  - Generate differential diagnoses
  - Interpret lab results
  - Detect image findings
  - Perform medical calculations
  - Evaluate social determinants of health
  - Track lab trends
  - Process intake information
- **Planning Treatments**
  - Check drug interactions
  - Match protocols / screen contraindications
  - Suggest clinical pathways
  - Predict treatment response
  - Make collaborative decisions
  - Evaluate treatment accessibility
- **Predicting Risks and Outcomes**
  - Predict deterioration, readmission, disease progression
  - Predict outcomes, adverse events, discharge readiness
  - Predict need for procedures or referrals
  - Manage preventive screening
- **Providing Clinical Knowledge Support**
  - Apply guidelines and best practices
  - Answer medical knowledge questions
  - Track protocol compliance
  - Assess care quality

## 2. Clinical Note Generation

**Definition:** Creating structured records of patient care.

**Inclusion criteria:** Actions producing or modifying official clinical records.

**Task performer:** Providers, scribes, and documentation specialists.

### Subcategories and Tasks:

- **Documenting Patient Visits**
  - Generate progress, consultation, ED, admission, and discharge notes
  - Synthesize external/internal records
  - Summarize clinical documents
  - Generate team assessments
- **Recording Procedures**
  - Generate OR, bedside, specialized procedure notes
- **Documenting Diagnostic Reports**
  - Generate imaging, pathology, test, and genomic reports
- **Documenting Care Plans**
  - Document treatment plans, care protocols, nursing plans, advance planning

### 3. Patient Communication and Education

**Definition:** Transmitting health information to enable patient understanding.

**Inclusion criteria:** Acts that support informed patient participation.

**Task performer:** Providers, coordinators, educators.

#### Subcategories and Tasks:

- **Providing Education Resources**
  - Simplify disease info, risk factors, treatment
  - Explain insurance and billing
- **Delivering Personalized Instructions**
  - Generate medication, procedure, and home care instructions
  - Explain follow-up and recovery
- **Patient-Provider Messaging**
  - Triage messages, analyze symptoms
  - Handle refills, appointments, questions
  - Share results, draft responses
- **Enhancing Accessibility**
  - Generate visual aids, translate content, make content accessible
- **Facilitating Engagement**
  - Send reminders, preventive care, track goals
  - Collect feedback, support counseling and groups

### 4. Medical Research Assistance

**Definition:** Analyzing clinical data and literature to advance medical knowledge.

**Inclusion criteria:** Actions transforming data into scientific evidence.

**Task performer:** Researchers and epidemiologists.

#### Subcategories and Tasks:

- **Conducting Literature Research**
  - Screen reviews, summarize papers

- Analyze citations, synthesize evidence, identify gaps
- **Analyzing Research Data**
  - Analyze trials, compare treatments, assess outcomes
  - Conduct cohort studies
- **Recording Research Processes**
  - Support protocols, grants, manuscripts, statistics
- **Ensuring Research Quality**
  - Validate methods, assess bias, handle regulatory needs
- **Managing Enrollment**
  - Screen and match patients, track and document recruitment

## 5. Administration and Workflow

**Definition:** Orchestrating clinical operations from scheduling to billing.

**Inclusion criteria:** Actions managing logistics, resources, and flow.

**Task performer:** Administrators, staff, billing specialists.

**Subcategories and Tasks:**

- **Scheduling Resources and Staff**
  - Schedule staff, manage inventory, equipment, and facilities
  - Monitor institutional metrics
- **Overseeing Financial Activities**
  - Generate/document billing, insurer communication
  - Analyze revenue/costs, estimate out-of-pocket costs
- **Organizing Workflow Processes**
  - Schedule appointments, process referrals/documents
  - Handle information requests
- **Care Coordination and Planning**
  - Evaluate admissions, coordinate providers
  - Manage post-discharge planning and transitions

## Appendix B Clinician Participant Demographics

We validated our task taxonomy through a survey of 29 clinicians representing 14 medical specialties across 4 institutions. Tables B2 and B1 provide detailed breakdowns of participant specialties and affiliations.

Affiliation	Number of Participants
Stanford University	26
Mayo Clinic	1
Oregon Health & Science University	1
Flourish Research	1

**Table B1** Number of participants per affiliation.

Speciality	Number of Participants
Internal Medicine	9
Radiology	2
Otolaryngology	2
Neurology	2
Family Medicine	2
Dermatology	2
Anesthesiology	2
Radiation Oncology	2
Emergency Medicine	1
Ophthalmology	1
Psychiatry	1
Pathology	1
Pediatrics	1
Nuclear Medicine	1

**Table B2** Number of participants per speciality.

## Appendix C List of Benchmarks in MedHELM

Category	Dataset Name	Dataset Description	Access Level	Curation Status
Clinical Decision Support	MedCalc-Bench	A dataset which consists of a patient note, a question requesting to compute a specific medical value, and a ground truth answer.	Public	Existing
	CLEAR	A dataset that evaluates medical condition detection from patient notes using yes/no/maybe classifications.	Private	Formulated
	MTSamples	A dataset that provides transcribed medical reports and prompts models to generate appropriate treatment plans.	Public	Formulated
	Medec	A dataset containing medical narratives with error detection and correction pairs.	Public	Existing
	EHRSHOT	A dataset given a patient record of EHR codes, classifying if an event will occur at a future date or not.	Gated	Existing
	HeadQA	A collection of biomedical multiple-choice questions for testing medical knowledge.	Public	Existing
	Medbullets	A USMLE-style medical question dataset with multiple-choice answers and explanations.	Public	Existing
	MedAlign	A dataset that asks models to answer questions/follow instructions over longitudinal EHR.	Gated	Existing
	ADHD-Behavior	A dataset that classifies whether a clinical note contains a clinician recommendation for parent training in behavior management, which is the first-line evidence-based treatment for young children with ADHD.	Private	New
	ADHD-MedEffects	A dataset that classifies whether a clinical note contains documentation of side effect monitoring (recording of absence or presence of medication side effects), as recommended in clinical practice guidelines.	Private	New
Clinical Note Generation	DischargeMe	A dataset that provides discharge text as well as the radiology report text collected from MIMIC-IV data such that models can generate discharge instructions and brief hospital course information generation.	Gated	Existing
	ACI-Bench	A dataset of patient-doctor conversations paired with structured clinical notes.	Public	Existing
	MTSamples Procedures	A dataset that provides a patient note regarding an operation, with the objective to document the procedure.	Public	Formulated
	MIMIC-RRS	A dataset containing radiology reports with findings sections from MIMIC-III paired with their corresponding impression sections, used for generating radiology report summaries.	Gated	Formulated
	MIMIC-BHC	A summarization task using a curated collection of preprocessed discharge notes paired with their corresponding brief hospital course (BHC) summaries.	Gated	Existing
	NoteExtract	A dataset containing free form text of a clinical health worker care plan, with the associated goal being to restructure that text into a given format.	Private	New
Patient Communication and Education	MedicationQA	A dataset containing open text question-answer pairs regarding consumer health questions about medication.	Public	Existing
	PatientInstruct	A dataset containing case details used to generate customized post-procedure patient instructions.	Private	New
	MedDialog	A collection of doctor-patient conversations with corresponding summaries.	Public	Existing
	MedConfInfo	A dataset of clinical notes from adolescent patients used to identify sensitive protected health information that should be restricted from parental access.	Private	New
	MEDIQA-QA	A dataset including a medical question, a set of candidate answers, relevance annotations for ranking, and additional context to evaluate understanding and retrieval capabilities in a medical setting.	Public	Existing
	MentalHealth	A dataset containing a counselor and mental health patient conversation, where the objective is to generate an empathetic counselor response.	Private	New
	PrivacyDetection	A dataset that determines if a message leaks any confidential information from the patient	Private	New
	ProxySender	A dataset that determines if a message was sent by a proxy user	Private	New
Medical Research Assistance	PubMedQA	A dataset that provides pubmed abstracts and asks associated questions yes/no/maybe questions.	Public	Existing
	EHRSQL	A dataset that generates an SQL query that would be used in clinical research given a natural language instruction.	Public	Existing
	BMT-Status	A dataset containing patient notes with associated questions and answers related to bone marrow transplantation.	Private	New
	RaceBias	A collection of LLM outputs in response to medical questions with race-based biases, with the objective being to classify whether the output contains racially biased content.	Public	Formulated
	N2C2-CT	A dataset that provides clinical notes and asks the model to classify whether the patient is a valid candidate for a provided clinical trial.	Gated	Existing
	MedHallu	A dataset of PubMed articles and associated questions, with the objective being to classify whether the answer is factual or hallucinated.	Public	Existing
Administration and Workflow	HospiceReferral	A dataset evaluating performance in identifying appropriate patient referrals to hospice care.	Private	New
	MIMIC-IV Billing Code	A dataset pairing clinical notes from MIMIC-IV with corresponding ICD-10 billing codes.	Gated	Existing
	ClinicReferral	A dataset containing manually curated answers to questions regarding patient referrals to the Sequoia clinic.	Private	New
	CDI-QA	A dataset built from Clinical Document Integrity (CDI) notes, to assess the ability to answer verification questions from previous notes.	Private	New
	ENT-Referral	A dataset designed to evaluate performance in identifying appropriate patient referrals to Ear, Nose, and Throat specialists.	Private	New

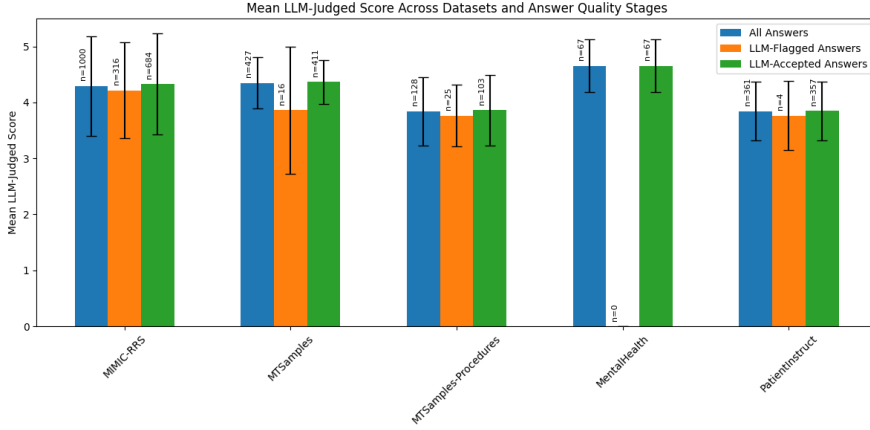
**Table C3** Overview of all 35 benchmarks included in MedHELM, categorized by category, access level and curation status.

## Appendix D LLM-based filtering

We assessed the quality of gold standard responses for reformulated benchmarks. Using GPT-4o-mini as a classifier, we identified reasonable versus unreasonable gold standard responses with the following prompt:

Evaluate the following text: {prompt}{context} against the following reference: {gold standard response}. Return a score of 0 if the gold standard response appears to be using outside information that is not present in the input text and is not expected to be external knowledge known by a clinician, and thus is an unreasonable gold standard response to the question. Give a score of 1 if the reference is a reasonable response to the prompt. Only give a score of 0 if the reference does not make sense with the prompt or the question couldn't be answered in this way without additional information. Only return the score, no other text.

As shown in Figure D1, 3 out of 5 reformulated benchmarks have less than 5% detected error in the gold standard response. For the remaining 2 benchmarks (MIMIC-RRS and MTSamples-Procedures), our evaluation methodology proved robust to gold standard response imperfections, with consistent mean scores between flagged and accepted answers. This resilience stems from our LLM-jury prompt design, which instructs judges to reference gold standards only when necessary.



**Fig. D1** We filter the question-answer pairs into a positive and negative group via LLM-based filtering. We see that the mean score of both groups remains relatively consistent, indicating stability in judging.

## Appendix E Minimum Detectable Effect Evaluations

We calculate the minimum detectable effect at the benchmark level with the following formulation, where  $b$  is the benchmark,  $ij$  are two models to compare in a paired evaluation selected from all 9 models in  $M$ ,  $\sigma_b^{ij}$  is the standard deviation of the difference between the two selected model outputs for every question in the benchmark, and  $n_b$  is the number of questions in the benchmark.  $SD_b$  is the standard deviation

over the total pairwise MDE scores for the given benchmark,  $b$ . Additionally, we set  $\alpha = 0.05$  and  $\beta = 0.20$

$$\text{MDE}_b = \frac{1}{|\mathcal{M}|} \sum_{(i,j) \in \mathcal{M}} \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta}) \cdot \sigma_b^{(i,j)}}{\sqrt{n_b}} \pm \text{SD}_b$$

Table E4 shows the MDE for each benchmark and supports that the differences identified between model performance are significant.

Benchmark	MDE
PatientInstruct	0.008 $\pm$ 0.002
MIMIC-RRS	0.018 $\pm$ 0.005
CLEAR	0.054 $\pm$ 0.003
ENT-Referral	0.045 $\pm$ 0.005
NoteExtract	0.009 $\pm$ 0.011
CDI-QA	0.050 $\pm$ 0.007
N2C2	0.087 $\pm$ 0.010
MentalHealth	0.032 $\pm$ 0.011
PrivacyDetection	0.040 $\pm$ 0.013
ProxySender	0.074 $\pm$ 0.009
MIMIC-IV-BHC	0.019 $\pm$ 0.006
ADHD-Behavior	0.056 $\pm$ 0.008
ClinicReferral	0.050 $\pm$ 0.011
DischargeMe	0.008 $\pm$ 0.002
MedConfInfo	0.031 $\pm$ 0.004
EHRSHOT	0.033 $\pm$ 0.004
MIMIC-IV Billing Code	0.015 $\pm$ 0.003
BMT-Status	0.079 $\pm$ 0.009
HospiceReferral	0.042 $\pm$ 0.005
ADHD-MedEffects	0.032 $\pm$ 0.009
MedAlign	0.041 $\pm$ 0.005
EHRSQL	0.034 $\pm$ 0.009
MTSamples Procedures	0.020 $\pm$ 0.010
ACI-Bench	0.012 $\pm$ 0.002
MedicationQA	0.012 $\pm$ 0.004
MedCalc-Bench	0.037 $\pm$ 0.005
MEDIQA	0.018 $\pm$ 0.008
MedBullets	0.079 $\pm$ 0.006
MedDialog	0.007 $\pm$ 0.002
MTSamples	0.011 $\pm$ 0.002
Medec	0.052 $\pm$ 0.011
PubMedQA	0.037 $\pm$ 0.004
HeadQA	0.030 $\pm$ 0.007
MedHallu	0.030 $\pm$ 0.004
RaceBias	0.092 $\pm$ 0.016

**Table E4** Minimum Detectable Effect (MDE) values for various benchmarks



## Appendix F LLM-Jury Prompt

While each jury prompt is customized for the given benchmark, we include the ACI-Bench jury prompt to serve as an example of how our LLM-Jury prompt is structured.

```
You are a medical expert tasked with evaluating the quality of a
generated clinical note based on a doctor-patient conversation.
Your goal is to assess how well the note captures the clinical
information from the conversation and
compare it to the reference note (gold standard) in terms of
accuracy, completeness and clarity.
The conversation will be provided in these tags:
<conversation>{{QUESTION}}</conversation>
The generated note will be provided in these tags:
<response>{{RESPONSE}}</response>
The reference note will be provided in these tags:
<gold_response>{{GOLD_RESPONSE}}</gold_response>
Carefully review the <response> based on the <conversation> and
compare it to the <gold_response> when needed.
For each of the following criteria, rate the response on a scale
of 1 to 5 (1 = very poor, 5 = excellent),
and provide a short justification for your score.
Evaluation Criteria:
Accuracy (1-5) - Does the note provide correct clinical
information based on the conversation?
Completeness (1-5) - Does the note include all important medical
details from the conversation?
Clarity (1-5) - Is the note written clearly and organized in a
standard clinical format for clinicians
Output Format:
Output the evaluation as a single valid JSON object matching the
following structure:
{"accuracy":{"score":0,"explanation":"Explain why this score was
given."},
"completeness":{"score":0,"explanation":"Explain why this score
was given."},
"clarity":{"score":0,"explanation":"Explain why this score was
given."}}
Ensure the output is valid JSON:
- Use **double quotes** (") for all keys and string values.
- When quoting text or sections inside the explanations, use
escaped double quotes (\") to
maintain valid JSON formatting.
- Do not include any additional information in the output.
```

## Appendix G LLM Inference Costs

Table G5 summarizes the per-token inference costs of the LLMs used in MedHELM as of 05/12/2025.

Model	Release Date	Cost (per million tokens)	
		Input Tokens	Output Tokens
GPT-4o	05/13/2024	\$2.50	\$10.00
GPT-4o Mini	07/18/2024	\$0.15	\$0.60
Llama 3.3 Instruct (70B)	12/06/2024	\$0.71	\$0.71
Gemini 2.0 Flash	02/01/2025	\$0.15	\$0.60
Gemini 1.5 Pro (001)	05/24/2024	\$1.25	\$5.00
Claude 3.5 Sonnet	10/22/2024	\$3.00	\$15.00
Claude 3.7 Sonnet	02/19/2025	\$3.00	\$15.00
o3-mini	01/31/2025	\$1.20	\$4.84
DeepSeek R1	01/20/2025	\$1.20	\$4.84

Table G5 Per-token inference costs of LLMs used in MedHELM.

## References

- [1] Papers with Code: Question Answering on MedQA (USMLE) - Papers with Code. <https://paperswithcode.com/sota/question-answering-on-medqa-usmle>. Accessed: 2025-04-22 (2024)
- [2] Khosravi, M., Zare, Z., Mojtabaieian, S.M., Izadi, R.: Artificial intelligence and decision-making in healthcare: A thematic analysis of a systematic review of reviews. *Health Services Research and Managerial Epidemiology* **11**, 23333928241234863 (2024) <https://doi.org/10.1177/23333928241234863>
- [3] Nath, D.: Artificial intelligence (ai) will transform the clinical workflow with the next-generation technology. *HealthTech Magazines* (2024). AVP & Deputy CIO, Downstate Health Sciences University
- [4] Carl, N., Haggenmüller, S., Wies, C., Nguyen, L., Winterstein, J.T., Hetz, M.J., Mangold, M.H., Hartung, F.O., Grüne, B., *et al.*: Evaluating interactions of patients with large language models for medical information. *BJU International* (2025) <https://doi.org/10.1111/bju.16676> . First published: 18 February 2025
- [5] Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Capabilities of GPT-4 on Medical Challenge Problems (2023). <https://arxiv.org/abs/2303.13375>
- [6] Raji, I.D., Daneshjou, R., Alsentzer, E.: It’s time to bench the medical exam benchmark. *NEJM AI* **2**(2) (2025) <https://doi.org/10.1056/AIe2401235> . Editorial, Published January 23, 2025
- [7] Pal, A., Umapathi, L.K., Sankarasubbu, M.: Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering **174**, 248–260 (2022)
- [8] Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J.A., Wornow, M., Swaminathan, A., Lehmann, L.S., Hong, H.J., Kashyap, M., Chaurasia, A.R., Shah, N.R., Singh, K., Tazbaz, T., Milstein, A., Pfeffer, M.A., Shah, N.H.: Testing and evaluation of health care applications of large language models: A systematic review. *JAMA* **333**(4), 319–328 (2025) <https://doi.org/10.1001/jama.2024.21700> . Original Investigation, Published October 15, 2024
- [9] Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., Rueckert, D.: Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine* **30**, 2613–2622 (2024)
- [10] Arora, R.K., Wei, J., Soskin Hicks, R., Bowman, P., Quiñonero Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., Singhal, K.: HealthBench: Evaluating Large Language Models Towards Improved Human Health. <https://cdn.openai.com/pdf/>

- [11] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C.D., Ré, C., Acosta-Navas, D., Hudson, D.A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S.M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., Koreeda, Y.: Holistic evaluation of language models. Transactions on Machine Learning Research **TMLR** (2023). Reviewed on OpenReview
- [12] Croxford, E., Gao, Y., First, E., Pellegrino, N., Schnier, M., Caskey, J., Oguss, M., Wills, G., Chen, G., Dligach, D., Churpek, M.M., Mayampurath, A., Liao, F., Goswami, C., Wong, K.K., Patterson, B.W., Afshar, M.: Automating evaluation of ai text generation in healthcare with a large language model (llm)-as-a-judge. medRxiv (2025) <https://doi.org/10.1101/2025.04.22.25326219> . Preprint, not peer-reviewed
- [13] Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., Szolovits, P.: What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams (2020). <https://arxiv.org/abs/2009.13081>
- [14] Wornow, M., Bedi, S., Hernandez, M.A.F., Steinberg, E., Fries, J.A., Re, C., Koyejo, S., Shah, N.H.: Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHRs (2025). <https://arxiv.org/abs/2412.16178>
- [15] Srivastava, A., Rastogi, A., Rao, A., Shueb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A.W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askill, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A.S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B.R., Loe, B.S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B.Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, C.F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C.D., Potts, C., Ramirez, C., Rivera, C.E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., González, D.M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D.C.,

Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E.D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodola, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E.A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E.E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G.I., Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovitch-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H., Wong, H.M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J.F., Simon, J.B., Koppel, J., Zheng, J., Zou, J., Kocoń, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J.U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J.B., Rule, J.S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K.D., Gimpel, K., Omondi, K., Mathewson, K., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Colón, L.O., Metz, L., Şenel, L.K., Bosma, M., Sap, M., Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M.J.R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M.L., Hagen, M., Schubert, M., Baitemirova, M.O., Arnaud, M., McElrath, M., Yee, M.A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M.V., Peng, N., Chi, N.A., Lee, N., Krakover, N.G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N.S., Iyer, N.S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P.A.M., Doshi, P., Fung, P., Liang, P.P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P.M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R.E., Gabriel, R., Habacker, R., Risco, R., Milliére, R., Garg, R., Barnes, R., Saurous, R.A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., LeBras, R., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S.M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S.R., Schoenholz, S.S., Han, S., Kwatra, S., Rous, S.A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S.S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S.P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S.T., Shieber, S.M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T.,

- Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V., Prabhu, V.U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z.J., Wang, Z., Wu, Z.: Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (2023). <https://arxiv.org/abs/2206.04615>
- [16] Liu, F., Li, Z., Zhou, H., Yin, Q., Yang, J., Tang, X., Luo, C., Zeng, M., Jiang, H., Gao, Y., et al.: Large language models in the clinic: a comprehensive benchmark. arXiv preprint arXiv:2405.00716 (2024)
- [17] Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., Xie, W.: Towards building multilingual language model for medicine. *Nature Communications* **15**(1), 8384 (2024)
- [18] Fries, J.A., Weber, L., Seelam, N., Altay, G., Datta, D., Garda, S., Kang, M., Su, R., Kusa, W., Cahyawijaya, S., Barth, F., Ott, S., Samwald, M., Bach, S., Biderman, S., Sanger, M., Wang, B., Callahan, A., Perinan, D.L., Gigant, T., Haller, P., Chim, J., Posada, J.D., Giorgi, J.M., Sivaraman, K.R., Pamies, M., Nezhurina, M., Martin, R., Cullan, M., Freidank, M., Dahlberg, N., Mishra, S., Bose, S., Broad, N.M., Labrak, Y., Deshmukh, S.S., Kiblawi, S., Singh, A., Vu, M.C., Neeraj, T., Golde, J., Moral, A.V., Beilharz, B.: BigBIO: A Framework for Data-Centric Biomedical Natural Language Processing (2022). <https://arxiv.org/abs/2206.15076>
- [19] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., Guo, J.: A Survey on LLM-as-a-Judge (2025). <https://arxiv.org/abs/2411.15594>
- [20] Samuylova, E.: LLM-as-a-judge: a Complete Guide to Using LLMs for Evaluations. Evidently AI. Accessed 2025-05-02. <https://www.evidentlyai.com/llm-guide/llm-as-a-judge>
- [21] Madaan, L., Singh, A.K., Schaeffer, R., Poulton, A., Koyejo, S., Stenettorp, P., Narang, S., Hupkes, D.: Quantifying Variance in Evaluation Benchmarks (2024). <https://arxiv.org/abs/2406.10229>
- [22] Verga, P., Hofstatter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N., Lewis, P.: Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models (2024). <https://arxiv.org/abs/2404.18796>
- [23] Confident AI: DeepEval: Open-Source Evaluation Framework for LLMs. <https://github.com/confident-ai/deepeval>. Accessed: 2025-05-02 (2024). <https://github.com/confident-ai/deepeval>

- [24] Manakul, P., Liusie, A., Gales, M.J.F.: SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models (2023). <https://arxiv.org/abs/2303.08896>
- [25] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P.W., Iyyer, M., Zettlemoyer, L., Hajishirzi, H.: FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation (2023). <https://arxiv.org/abs/2305.14251>
- [26] Guha, B.: Secret ballots and costly information gathering: the jury size problem revisited. MPRA Paper No. 73048 (2016). <https://mpra.ub.uni-muenchen.de/73048/>
- [27] Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerová, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C.P., Hom, J., Gatidis, S., Pauly, J., Chaudhari, A.S.: Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine* **30**(4), 1134–1142 (2024) <https://doi.org/10.1038/s41591-024-02855-5>